Solution 1: Kullback-Leibler Divergence

(a) Let f be the pmf of the $Bin(n,p)$ distribution and q the density of the $\mathcal{N}(\mu,\sigma^2)$.

(i)
$$D_{KL}(f||q) = \mathbb{E}_f[\log \frac{f(X)}{q(X,\theta)}] = \mathbb{E}_f[\log f(X)] - \mathbb{E}_f[\log q(X|\theta)]$$

(ii) For the gradients, we must derive the partial derivatives of the second part of the KLD. The involved log-density is
$$\log q(X|\theta) = const. - 0.5 \log \sigma^2 - \frac{1}{2\sigma^2}(X-\mu)^2.$$

$$\partial D_{KL}(f||q)/\partial\mu = \partial - \mathbb{E}_f \log[q(X|\theta)] = \mathbb{E}_f \frac{1}{\sigma^2}(X-\mu) \tag{1}$$

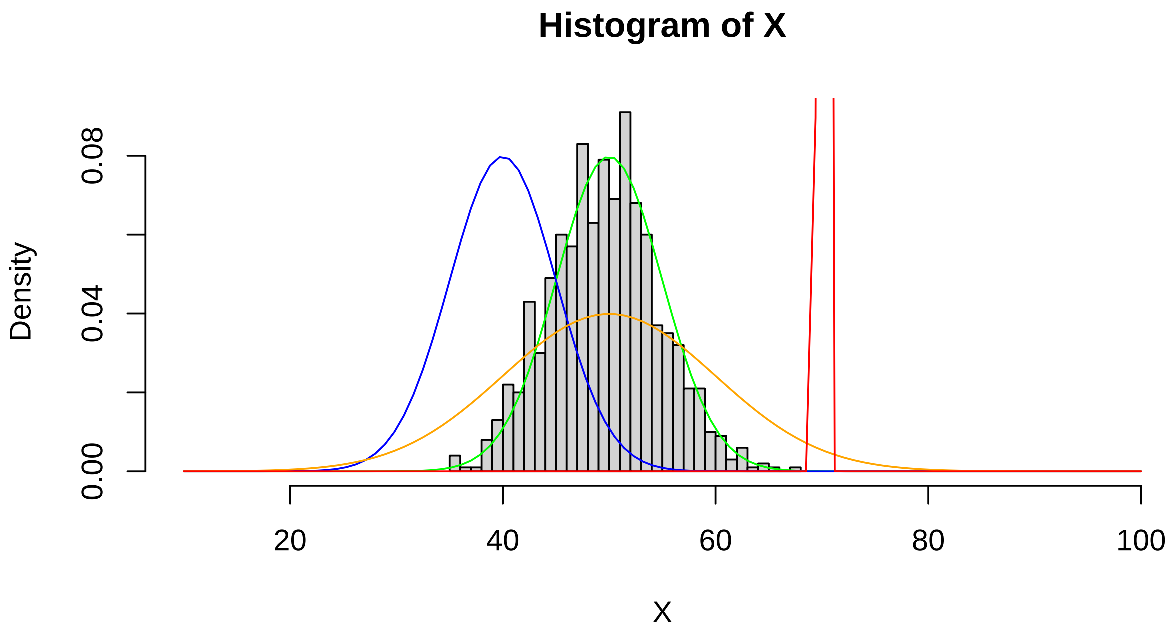
$$\partial D_{KL}(f||q)/\partial\sigma^2 = \partial - \mathbb{E}_f \log[q(X|\theta)] = \mathbb{E}_f[\frac{1}{2\sigma^2} + \frac{-1}{2\sigma^4}(X-\mu)^2] \tag{2}$$

(iii) Yes, there is. We can first set (1) to zero and get: $\mu = \mathbb{E}_f(X) \Leftrightarrow \mu = np$. We then use this solution for the second equation (2), which we also set to zero first:
$$(2) = 0 \Leftrightarrow \sigma^2 = \mathbb{E}_f[(X-\mu)^2] = \text{Var}_f(X) + (\mathbb{E}_f[X-\mu])^2 = np(1-p) + (\mathbb{E}_f[X-\mu])^2.$$

Using $\mu = np$, the second term vanishes and we get the optimal $\sigma^2 = np(1-p) = \text{Var}_f(X)$. Note that we would have to prove that the second derivative is < 0 to be sure that we found a minimum!

(iv) We could, alternatively, use the gradients and do gradient descent to find the optimal $\boldsymbol{\theta}$.

(b)
```r
nr_points = 1000
p = 0.5
n = 100
# create data
X <- rbinom(nr_points, prob = p, size = n)

# define different Normal density functions
normal_optimal <- function(x) dnorm(x, mean = n*p, sd = sqrt(n*p*(1-p)))
normal_shift <- function(x) dnorm(x, mean = n*p - 10, sd = sqrt(n*p*(1-p)))
normal_scale_increase <- function(x) dnorm(x, mean = n*p, sd = sqrt(n*p*(1-p))*2)
normal_right_scale_decrease <- function(x) dnorm(x, mean = n*p + 20, sd = p*(1-p))

hist(X, breaks = 25, xlim = c(10, 100), freq = FALSE)
curve(normal_optimal, from = 10, to = 100, add = TRUE, col = "green")
curve(normal_shift, from = 10, to = 100, add = TRUE, col = "blue")
curve(normal_scale_increase, from = 10, to = 100, add = TRUE, col = "orange")
curve(normal_right_scale_decrease, from = 10, to = 100, add = TRUE, col = "red")
```



For these distributions, we get the following KL divergence values (up to an additive constant):

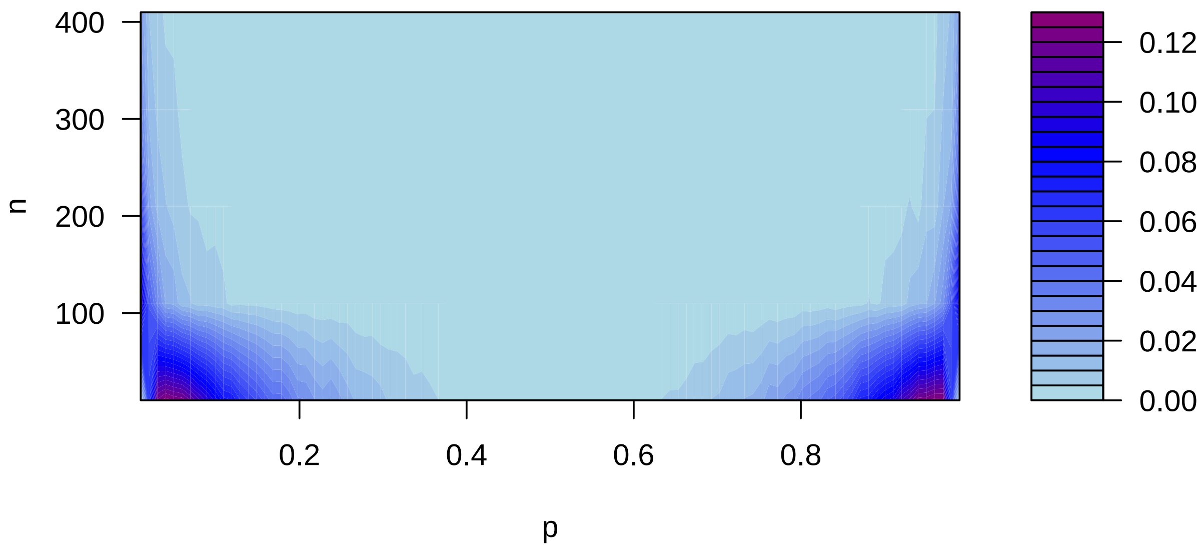
$$D_{KL}(f||q) = const. + 0.5 \log \sigma^2 + \frac{1}{2\sigma^2}(\mathrm{Var}_f(X) + (np - \mu)^2))$$

```
kld_value <- function(mu,sigma2)
{
  0.5*log(sigma2) +
    0.5 * (sigma2)^(-1) * (n*p*(1-p) + (n*p - mu)^2)
}
(optimal_green <- kld_value(n*p,n*p*(1-p)))

## [1] 2.109438

(shift_blue <- kld_value(n*p-10,n*p*(1-p)))

## [1] 4.109438

(scale_increase_orange <- kld_value(n*p,n*p*(1-p)*4))

## [1] 2.427585

(right_scale_decrease_red <- kld_value(n*p+20, (p*(1-p))^2))

## [1] 3398.614
```

(c) Since we are now required to calculate the exact KLD values, we would also have to calculate $\mathbb{E}_f(\log f(X))$, which is somewhat more difficult. If you search the internet for a solution ($\rightarrow$ "entropy of a binomial distribution"), you will find an approximate solution using the de-Moivre-Laplace theorem. Alternatively, we could make use of the central limit theorem, but then we would just approximate f with a normal distribution with $\mu = np$ and $\sigma^2 = np(1-p)$, which would give us a constant KLD of zero (the very same happens if you use the first approximation using the de-Moivre-Laplace-theorem). We here instead will approximate the expectation using a large sample from the true underlying distribution:

$$D_{KL}(f||q) \approx \frac{1}{B}\sum_{b=1}^{B}[\log f(X) - \log q(X|\mu = np, \sigma^2 = np(1-p))]$$

```
p_seq <- seq(0.01, 0.99, l = 100)
n_seq <- seq(10, 500, by = 100)
B <- 10000

kld_value_approx <- function(n,p){

  # sample a large number of data points from true distribution
  x <- rbinom(B, prob = p, size = n)

  # approximate the mean; threshold values to 0 if < 0 due
  # to the approximation
  pmax(
    mean(
      dbinom(x, prob = p, size = n, log = TRUE) -
        dnorm(x, mean = n*p, sd = sqrt(n*p*(1-p)), log = TRUE),
      na.rm = TRUE
    ),
    0)

}

kld_val <- sapply(n_seq, function(this_n)
  sapply(p_seq, function(this_p) kld_value_approx(this_n, this_p)))

cols = rev(colorRampPalette(c('darkred','red','blue','lightblue'))(50))

filled.contour(x = p_seq, y = n_seq, z = kld_val,
               xlab = "p", ylab = "n",
               col = cols
               )
```



(d) Based on the previous result, one can see that the KLD is very close to zero but has larger values for very small or very large values of p and / in combination with a small number of experiments n. These are exactly the cases where the normal approximation of a binomial distribution does not work so well.

Solution 2: The Convexity of KL Divergence

(a) We expand the left side of the inequality and obtain:

$$\begin{aligned}
&D_{KL}(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \\
&= \int_{\mathcal{X}} \left((\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \right) \mathrm{d}x \\
&\leq \int_{\mathcal{X}} \left(\lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} \right) \mathrm{d}x \\
&= \lambda \int_{\mathcal{X}} \left(p_1(x) \log \frac{p_1(x)}{q_1(x)} \right) \mathrm{d}x + (1-\lambda) \int_{\mathcal{X}} \left(p_2(x) \log \frac{p_2(x)}{q_2(x)} \right) \mathrm{d}x \\
&= \lambda D_{KL}(p_1||q_1) + (1-\lambda) D_{KL}(p_2||q_2).
\end{aligned} \tag{3}$$